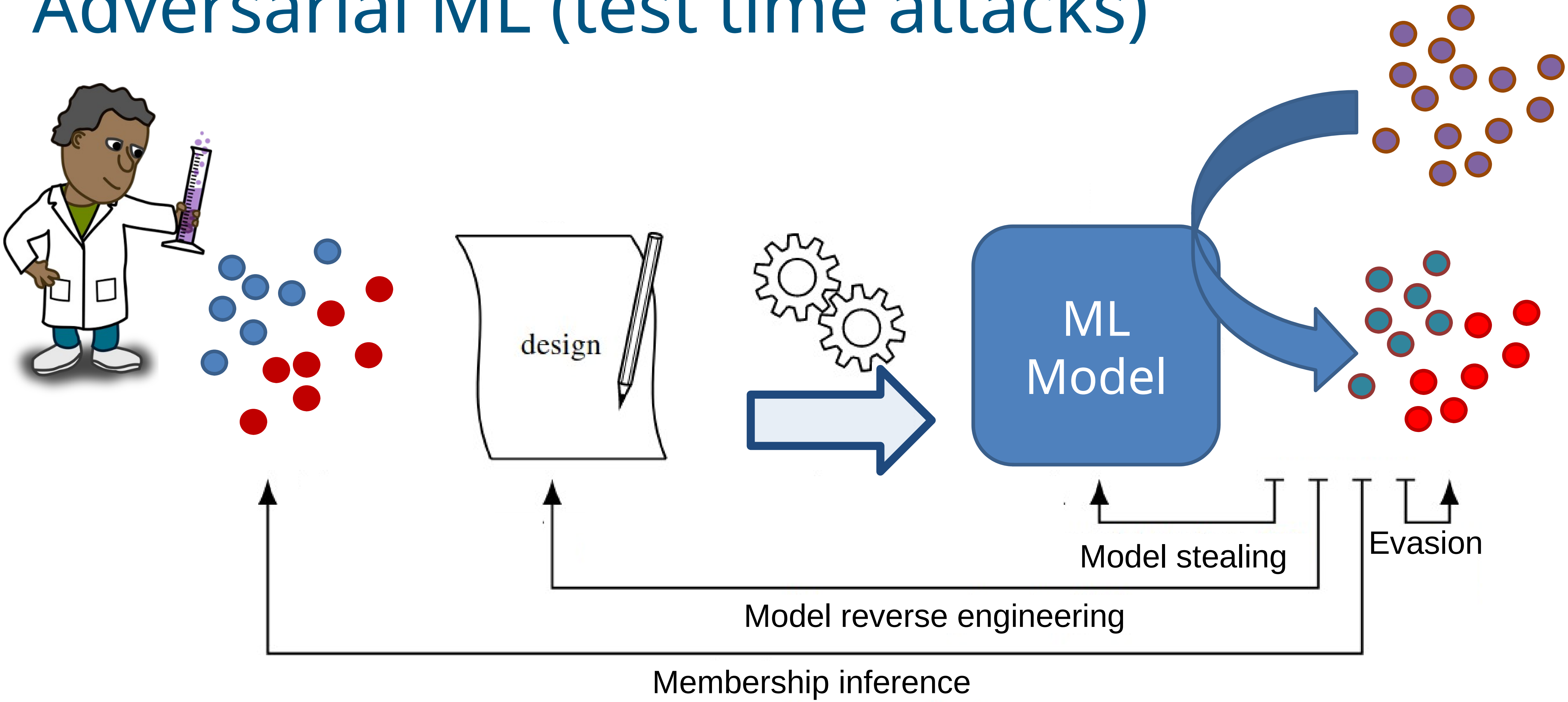
CISPA
HELMHOLTZ CENTER FOR
INFORMATION SECURITY

Killing four Birds with one Gaussian Process: The relation between different Test-time attacks

Adversarial ML (test time attacks)



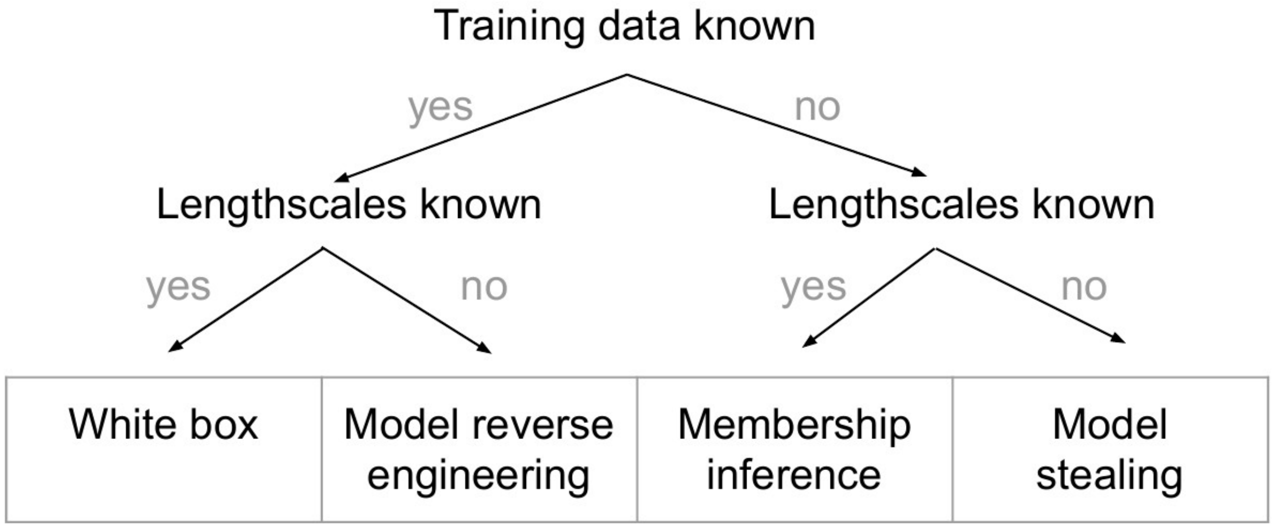
Why Gaussian Processes?

GP, after training, **are fully specified** and **deterministic**

GP’s **curvature** can be set using the **lengthscale**

GP are applied in **medical** settings, risk assessment is **crucial**

GP allow to **relate** IP based attacks:



Training data known

yes

no

Lengthscales known

Lengthscales known

yes

no

yes

no

White box

Model reverse engineering

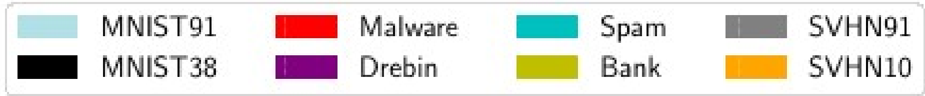
Membership inference

Model stealing

Goal of this paper

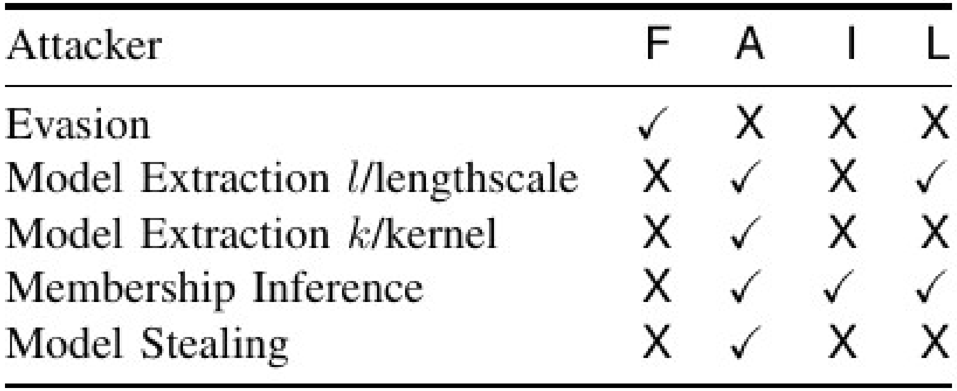
Study test-time attacks **in relation**, **not individually**

Use a **range** of **different** Data-sets.



🟦	MNIST91	🟥	Malware	🟦	Spam	⬛	SVHN91
⬛	MNIST38	🟪	Drebin	🟨	Bank	🟧	SVHN10

Threat models using FAIL*:



Attacker	F	A	I	L
Evasion	✓	X	X	X
Model Extraction l/lengthscale	X	✓	X	✓
Model Extraction k/kernel	X	✓	X	X
Membership Inference	X	✓	✓	✓
Model Stealing	X	✓	X	X

*Suciu et al., USENIX 2018

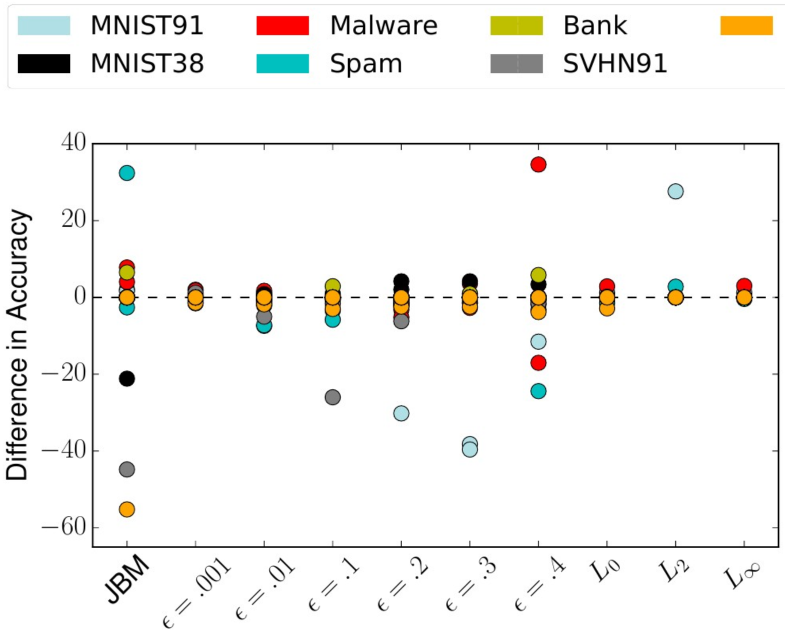
Evasion

Test **transferred** adversarial examples from **DNN, SVM, GP**

Compare long and short lengthscales (**steep** and **low curvature**)

Steep curvature is **harder** to attack with **one step attacks**

Low curvature is **harder** to attack with optimized **attacks**



Long lengthscale performs better

Short lengthscale (steep curvature) performs better

One step attacks

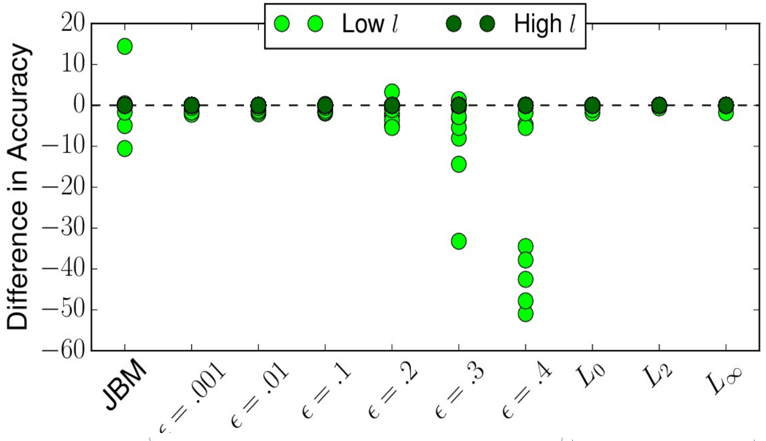
Optimized attacks

Evasion II - rejection

Test **transferred** adversarial examples from **DNN, SVM, GP**

Compare long and short lengthscales (**steep** and **low curvature**) and **reject** data if **output of GP is 0**

Only a classifier with **steep** curvature **benefits** from **rejection**



GP without rejection is better

GP with rejection (at µ=0.0) Performs better

One step attacks

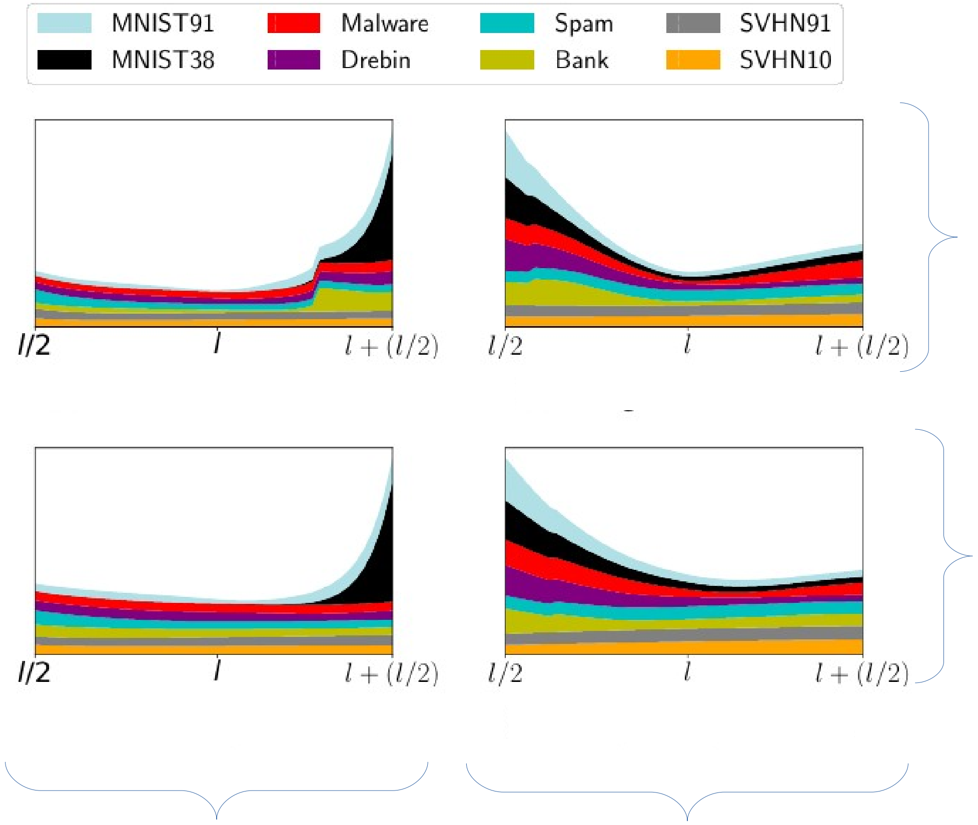
Optimized attacks

Model reverse engineering - lengthscale

Compare long and short lengthscales (**steep** and **low curvature**)

Try to infer **lengthscale** given **partial or no access** to used training data

A **short lengthscale** conceals the lengthscale **better**



Training data partially known

Training data unknown

short lengthscale

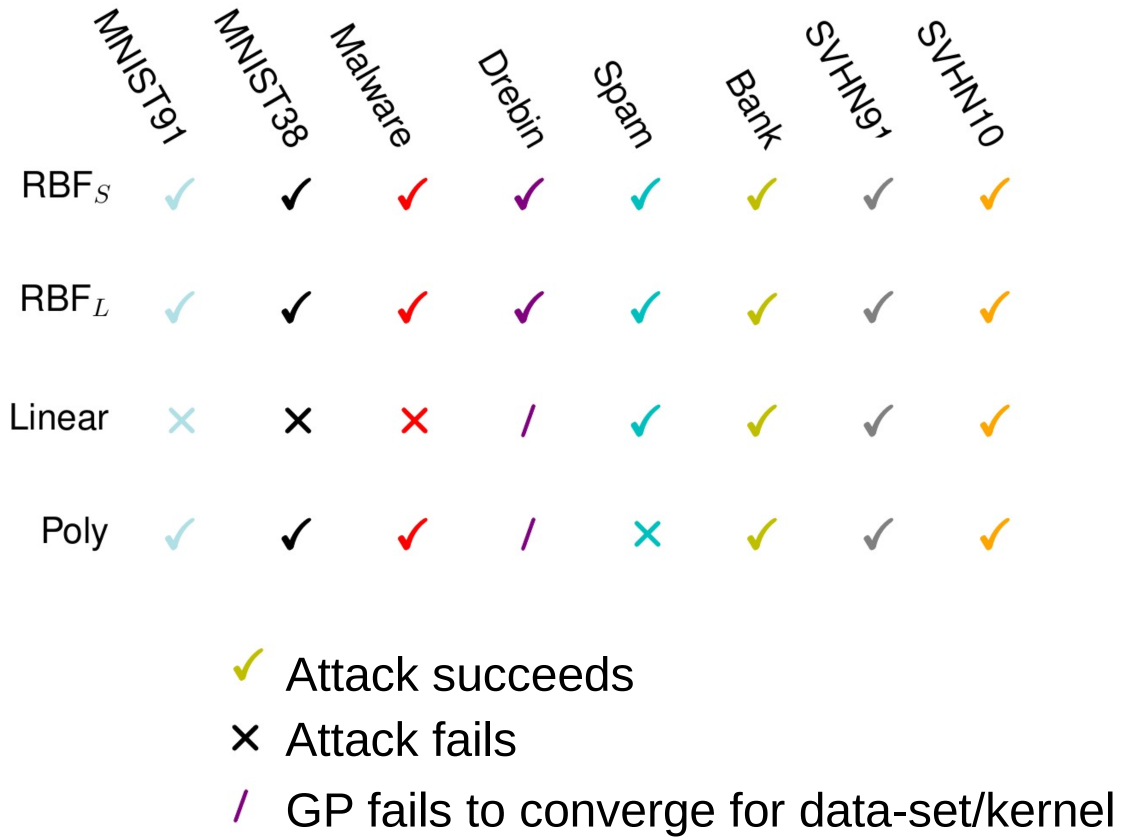
long lengthscale

Model reverse engineering - kernel

Compare long and short lengthscales (**steep** and **low curvature**)

Attempt to infer **kernel** used in GP

Attack is successful **regardless** of **curvature in RBF kernel**



	MNIST91	MNIST38	Malware	Drebin	Spam	Bank	SVHN91	SVHN10
RBF$_S$	✓	✓	✓	✓	✓	✓	✓	✓
RBF$_L$	✓	✓	✓	✓	✓	✓	✓	✓
Linear	×	×	×	/	✓	✓	✓	✓
Poly	✓	✓	✓	/	×	✓	✓	✓

✓ Attack succeeds

× Attack fails

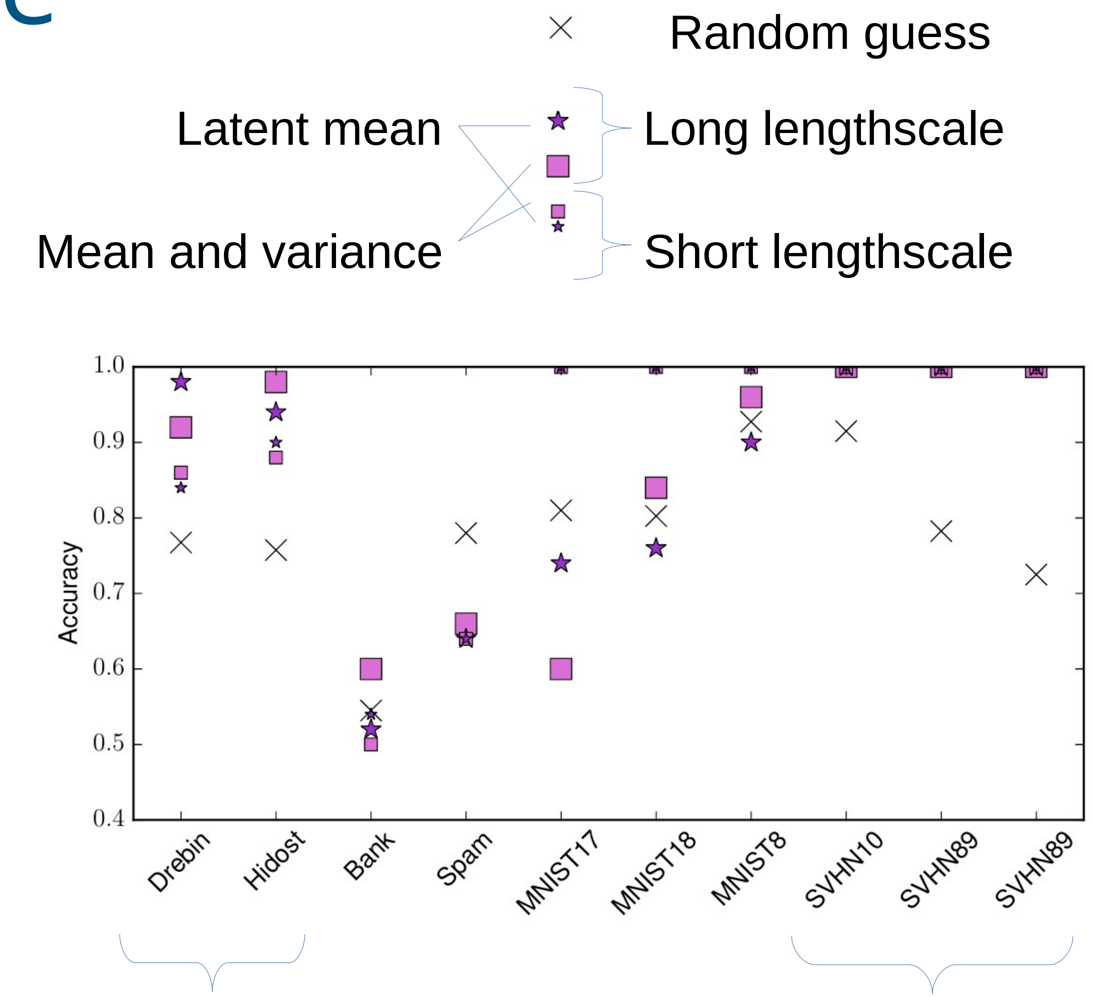
/ GP fails to converge for data-set/kernel

Membership inference

Compare long and short lengthscales (**steep** and **low curvature**)

Try to infer if point is **in training data** given **latent mean / mean and variance**

A long lengthscale **is more robust** towards **membership inference**



sparse
(fixable, see paper)

our experiments show
data-set shift for GP

Conclusion

AML attacks **should not** be studied **in isolation**.

Defending one attack might **increase** vulnerability for an **unrelated** attack!

A **short** lengthscale is harder to attack with optimized attacks

A **short** lengthscale conceals the lengthscale better

Attack is successful **regardless of curvature** in RBF kernel

A **long** lengthscale is more robust towards membership inference

Thank you!